

Is ChatGPT-5 Ready for Mammogram VQA?

Qiang Li¹ Shansong Wang¹ Mingzhe Hu¹ Mojtaba Safari¹ Zachary Eidex¹ Xiaofeng Yang¹✉

¹Department of Radiation Oncology, Winship Cancer Institute, Emory University School of Medicine

✉ Corresponding author: xiaofeng.yang@emory.edu

Abstract

Mammogram visual question answering (VQA) integrates image interpretation with clinical reasoning and has potential to support breast cancer screening. We systematically evaluated the GPT-5 family and GPT-4o model on four public mammography datasets (EMBED, InBreast, CMMD, CBIS-DDSM) for BI-RADS assessment, abnormality detection, and malignancy classification tasks. GPT-5 consistently was the best performing model but lagged behind both human experts and domain-specific fine-tuned models. On EMBED, GPT-5 achieved the highest scores among GPT variants in density (56.8%), distortion (52.5%), mass (64.5%), calcification (63.5%), and malignancy (52.8%) classification. On InBreast, it attained 36.9% BI-RADS accuracy, 45.9% abnormality detection, and 35.0% malignancy classification. On CMMD, GPT-5 reached 32.3% abnormality detection and 55.0% malignancy accuracy. On CBIS-DDSM, it achieved 69.3% BI-RADS accuracy, 66.0% abnormality detection, and 58.2% malignancy accuracy. Compared with human expert estimations, GPT-5 exhibited lower sensitivity (63.5%) and specificity (52.3%). While GPT-5 exhibits promising capabilities for screening tasks, its performance remains insufficient for high-stakes clinical imaging applications without targeted domain adaptation and optimization. However, the tremendous improvements in performance from GPT-4o to GPT-5 show a promising trend in the potential for general large language models (LLMs) to assist with mammography VQA tasks.

1 Introduction

Breast cancer is the most frequently diagnosed malignancy among women worldwide and a leading cause of cancer-related death [1]. Mammography remains the cornerstone of large-scale breast cancer screening, providing a critical tool for identifying abnormal tissue changes. Interpreting these images, however, is a complex and time-consuming task that demands specialized expertise and extensive clinical experience. The variability in breast tissue patterns, combined with the often subtle presentation of early-stage cancers, makes accurate reading challenging [2]. Even among expert radiologists, performance can vary significantly [3]. These inherent difficulties have spurred the advancement of artificial intelligence technologies aimed at supporting radiologists, improving diagnostic accuracy, and promoting greater consistency in interpretations. [4, 5, 6, 7].

Given the complexity of medical image interpretation, zero-shot application of large language or vision language models without any domain-specific fine-tuned may yield suboptimal results[8, 9]. AI models fine-tuned on medical imaging data have demonstrated strong performance across multiple radiology modalities and downstream tasks, including classification, detection, and segmentation, often surpassing traditional CAD systems in sensitivity and specificity [10]. Large vision-language models further extend these advances, and when fine-tuned on domain-specific medical images, have demonstrated strong performance across several imaging modalities, including radiography, pathology, and retinal imaging. Medical visual question answering (VQA) builds on this foundation by combining image understanding with natural language reasoning, enabling models to answer clinically framed questions directly [11, 12, 13]. This paradigm offers a practical framework to test whether general-purpose multimodal models like ChatGPT-5 can interpret mammograms in ways that align with clinical decision-making.

Given recent reports of GPT-5's impressive performance gains over previous generations, we ask "Is ChatGPT-5 ready for mammogram VQA?" We evaluate the zero-shot performance of ChatGPT-5 with multimodal capabilities on mammogram interpretation tasks. Using a diverse set of U.S. screening mammograms, we assess its ability

to classify malignancy and related findings from single-view images. We construct VQA items from four public datasets: **EMBED** [14], **INBreast** [15], **CMMD** [16], and **CBIS-DDSM** [17], then tailoring the metadata and annotations into standardized templates. To control external factors, we use a fixed zero-shot prompt and evaluate each case on a single mammogram view without multi-view or clinical context.

2 Methodology

2.1 Datasets

To evaluate GPT-5 on digital X-ray mammography reasoning, we derive clinically relevant VQA items from the structured (panel) labels available in four public mammography datasets: EMBED [14], INBreast [15], CMMD [16], and CBIS-DDSM [17]. We harmonize dataset-specific metadata and clinical annotations (laterality, view, image type, BI-RADS breast density and assessment, lesion types/attributes, and biopsy-confirmed pathology) into a common schema and convert them into standardized question templates. These include lesion presence/type (mass, calcification, none, both), malignancy classification (benign vs. malignant), breast density (A–D), and ROI-conditioned attribute questions (e.g., mass shape/margins or calcification distribution).

- **EMBED** [14] contains 3.4 million screening and diagnostic images from 110,000 patients collected between 2013 and 2020 and focuses on generalizability across ethnic groups by providing a balanced representation of Black and White women [14]. It includes 2D, synthetic 2D (C-view), and 3D digital breast tomosynthesis (DBT) images, along with 60,000 annotated lesions linked to structured imaging descriptors and pathology-confirmed outcomes across six severity classes. In this study, we use the publicly available subset, which represents 20% of the 2D portion of the dataset.
- **InBreast** [15] comprises 115 cases (410 images), including 90 cases from women with bilateral breast involvement (four images per case) and 25 cases from mastectomy patients (two images per case). It encompasses a variety of lesion types, such as masses, calcifications, asymmetries, and distortions.
- **CMMD** [16] is a publicly available mammography dataset curated by the Cancer Hospital of the Chinese Academy of Medical Sciences [16]. It contains full-field digital mammography (FFDM) images from 1,775 patients collected between 2012 and 2016, accompanied by clinical information such as age, pathology results, and lesion type. Images are labeled at the breast level as benign or malignant based on pathology-confirmed outcomes. CMMD provides a valuable resource for developing and evaluating computer-aided diagnosis methods on a non-Western population.
- **CBIS-DDSM** [17] is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. Using CBIS-DDSM, we construct VQA items focused on lesion-type recognition (mass vs. calcification vs. none), ROI-aware localization/attributes, and pathology-linked questions where available. To ensure consistent evaluation coverage across datasets and question categories, we adopt a balanced sampling strategy: for each question type and answer category, we randomly sample an equal number of items, thereby avoiding bias toward over-represented classes in large datasets while keeping the total evaluation set size tractable.

2.2 VQA generation

The visual question answering (VQA) pairs were constructed from four publicly available mammography datasets: EMBED, InBreast, CMMD, and CBIS-DDSM. For each case, questions were automatically generated from the structured panel data and metadata accompanying the images, including patient information, imaging annotations, lesion type, BI-RADS density category, and biopsy-confirmed pathology. This design yields questions that target well-defined clinical tasks, such as “What is the BI-RADS breast density?” or “Is the imaging finding suggestive of malignancy?”, ensuring a one-to-one correspondence between the query, its answer, and the gold-standard label. By avoiding free-text questions based on subjective interpretation of image texture, morphology, or density patterns, the approach reduces variability and enhances reproducibility in automated evaluation. However, the absence of texture-oriented, descriptive queries limits the direct assessment of visual reasoning processes, positioning the evaluation as a test of structured clinical label inference rather than nuanced image-based reasoning.

2.3 Prompt Design

We evaluate GPT-5 using a zero-shot chain-of-thought (CoT) prompting strategy [18], implemented as a concise two-turn dialogue. In the first turn, a system message establishes the medical domain context, and the user presents the question, explicitly triggering step-by-step reasoning with the cue - **Let’s think step by step**. For multimodal

questions, all associated images are provided in this message as `image_url` entries, enabling the model to jointly process visual and textual information within a single reasoning phase. The model responds with an unconstrained explanatory rationale (*prediction_rationale*) without selecting a final answer.

In the second turn, the user issues a convergence instruction: **Therefore, among A through {END_LETTER}, the answer is** - where {END_LETTER} denotes the last available option label. At this stage, the model outputs only a single choice letter (*prediction*). This separation ensures that the reasoning process and final decision are elicited in distinct conversational steps. The jsonl templates for both text-only and image-augmented formats follow this protocol, using placeholders {QUESTION_TEXT}, {END_LETTER}, {IMAGE_URL_1}, {ASSISTANT_RATIONALE}, and {ASSISTANT_FINAL}. The prompting design template for the mammogram VQA task is illustrated in Fig. 1.

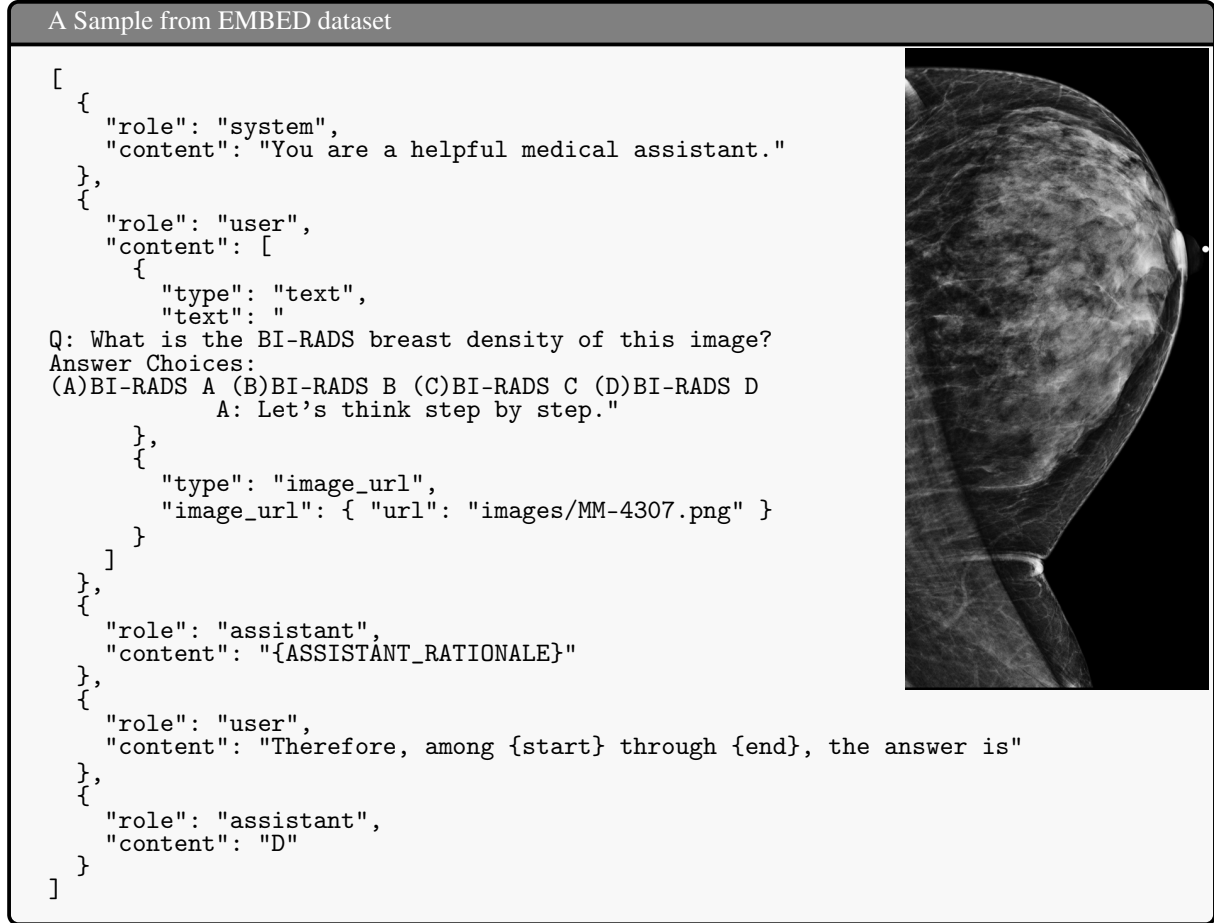


Figure 1: An example multimodal prompt from the EMBED dataset. The model is tasked with provided the correct answer given a mammogram and multiple choice question.

3 Results

3.1 Performance of GPTs on public mammograms dataset screening

Table 1-4 summarizes the screening performance (%) of GPT family models on multiple public mammography datasets, including EMBED, InBreast, CMMD, and CBIS-DDSM. Across datasets, GPT-5 consistently outperforms its smaller variants (GPT-5-mini, GPT-5-nano) and the previous-generation GPT-4o (GPT-4o-2024-11-20 version), particularly in tasks requiring fine-grained lesion characterization such as mass and calcification detection. While GPT-4o demonstrates reasonable accuracy in several settings, the performance gap relative to GPT-5 is more pronounced in multi-class tasks (e.g., BI-RADS classification) and malignancy assessment. Across four public mammography datasets (EMBED, InBreast, CMMD, CBIS-DDSM), GPT-5 shows measurable but still limited screening performance compared to specialized medical AI systems and human experts. On the EMBED dataset,

GPT-5 achieves the highest performance among GPT variants (e.g., Density: 56.8%, Distortion: 52.5%, Mass: 64.5%, Calcification: 63.5%, Malignancy: 52.8%), outperforming GPT-5-mini, GPT-5-nano, and GPT-4o, but still lagging behind pretrained fine-tuned SOTA models.

On the InBreast dataset, GPT-5 attains 36.9% BI-RADS accuracy, 45.9% for abnormality detection, and 35.0% for malignancy classification, far below the 90.6% malignancy accuracy of MRSN. In CMMD, GPT-5 records 32.3% abnormality detection and 55.0% malignancy accuracy, versus 79.7% from HybMNet.

On CBIS-DDSM, GPT-5 achieves 69.3% BI-RADS accuracy, 66.0% abnormality detection, and 58.2% malignancy accuracy. Compared to human expert estimations (sensitivity 86.9%, specificity 88.9%), GPT-5’s sensitivity drops to 63.5% ($\downarrow 23.4\%$) and specificity to 52.3% ($\downarrow 36.6\%$). Specialized models such as PHYSnet (82.0% malignancy accuracy) and ResNet18-S896 (79.6%) substantially outperform GPT-5.

Overall, GPT-5 consistently surpasses smaller GPT variants but remains notably behind both human experts and domain-optimized AI models, indicating substantial room for improvement in domain adaptation, reasoning transparency, and uncertainty calibration for high-stakes clinical imaging tasks.

These results highlight the potential of large-scale, instruction-tuned multimodal language models for zero-shot mammography interpretation, though specialized fine-tuned state-of-the-art (SOTA) models still achieve superior results on certain tasks, indicating room for further domain adaptation.

Table 1: Performance comparison between GPTs and pretrained fine-tuned SOTA models on the EMBED dataset.

| Model \ Tasks | Density | Distortion | Mass | Calcification | Malignancy |
|----------------------------|----------------|-------------------|-------------|----------------------|-------------------|
| GPT-5 | 56.8 | 52.5 | 64.5 | 63.5 | 52.8 |
| GPT-5-mini | 34.3 | 53.5 | 60.8 | 57.3 | 47.3 |
| GPT-5-nano | 24.8 | 53.5 | 52.3 | 51.5 | 47.8 |
| GPT-4o | 24.3 | 20.0 | 50.0 | 44.3 | 42.5 |
| Mammo-CLIP (ViT-B/16) [19] | - | - | - | - | 79.0 |
| Mammo-CLIP (ViT-L/14) [19] | - | - | - | - | 82.3 |

Table 2: Performance comparison between GPTs and pretrained fine-tuned SOTA models on the InBreast dataset.

| Model \ Tasks | BI-RADS | Abnormality calcification and mass | Malignancy |
|----------------------|----------------|---|-------------------|
| GPT-5 | 36.9 | 45.9 | 35.0 |
| GPT-5-mini | 28.1 | 49.2 | 40.0 |
| GPT-5-nano | 17.6 | 37.8 | 21.5 |
| GPT-4o | 23.7 | 36.1 | 30.4 |
| MRSN [20] | - | - | 90.6 |
| GGP [20] | - | - | 88.5 |

Table 3: Performance comparison between GPTs and pretrained fine-tuned SOTA models on the CMMD dataset.

| Model \ Tasks | BI-RADS | Abnormality calcification and mass | Malignancy |
|----------------------|----------------|---|-------------------|
| GPT-5 | - | 32.3 | 55.0 |
| GPT-5-mini | - | 42.3 | 63.3 |
| GPT-5-nano | - | 34.0 | 52.7 |
| GPT-4o | - | 38.5 | 48.5 |
| HybMNet [21] | - | - | 79.7 |
| GMIC [21] | - | - | 73.6 |

3.2 Comparison with human experts

Table 5 compares the malignant screening performance of different GPT model variants with human expert estimations on the CBIS-DDSM dataset. Human readers achieve a sensitivity of 86.9% and specificity of 88.9%,

Table 4: Performance comparison between GPTs and pretrained fine-tuned SOTA models on the CBIS-DDSM dataset.

| Model \ Tasks | BI-RADS | Abnormality calcification and mass | Malignancy |
|---------------------------------|---------|---------------------------------------|------------|
| GPT-5 | 69.3 | 66.0 | 58.2 |
| GPT-5-mini | 43.6 | 53.3 | 43.5 |
| GPT-5-nano | 20.2 | 41.8 | 39.0 |
| GPT-4o | 28.4 | 46.2 | 40.0 |
| MRSN [20] | - | - | 77.8 |
| Single view baseline CNN [22] | - | - | 71.0 |
| Single view evidential CNN [22] | - | - | 72.6 |
| PHYSnet [23] | - | - | 82.0 |
| ResNet18-S896 [24] | - | - | 79.6 |

whereas all GPT models fall substantially short in both metrics. GPT-5 attains the highest accuracy among model variants (58.2%) but still shows marked reductions in sensitivity (-23.4%) and specificity (-36.6%) relative to human performance. Lighter GPT-5 variants (mini and nano) and GPT-4o exhibit even larger performance gaps, with sensitivity drops exceeding 50% and specificity deficits up to 26.0%. These findings reflect the intrinsic challenges of mammogram interpretation, where diagnostic decisions hinge on subtle, low-contrast imaging features, and emphasize that without domain-specific fine-tuning, current general-purpose multimodal LLMs remain far from achieving expert-level screening accuracy.

Table 5: Comparison with Human Experts on CBIS-DDSM (Malignant Screening).

| | Sensitivity | Specificity | ACC |
|--------------------------|---------------------------------|---------------------------------|------|
| Human Estimation [2, 24] | 86.9 | 88.9 | – |
| GPT-5 | 63.5 ($\downarrow 23.4\%$) | 52.3 ($\downarrow 36.6\%$) | 58.2 |
| GPT-5-mini | 35.8 ($\downarrow 51.1\%$) | 62.9 ($\downarrow 26.0\%$) | 44.9 |
| GPT-5-nano | 34.9 ($\downarrow 52.0\%$) | 67.5 ($\downarrow 21.4\%$) | 51.2 |
| GPT-4o | 33.3 ($\downarrow 53.6\%$) | 66.7 ($\downarrow 22.2\%$) | 40.0 |

3.3 Case Studies on GPT Errors in Mammogram VQA

We selected four representative examples, two correctly classified and two misclassified, to qualitatively examine GPT-5’s decision-making in BI-RADS density assessment and malignancy detection tasks. These cases illustrate the model’s strengths in recognizing prototypical imaging features and its weaknesses in borderline density categorization and atypical lesion interpretation, shown in

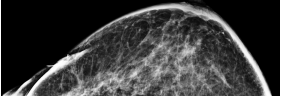
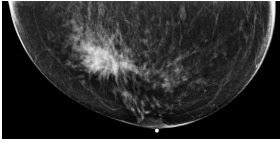
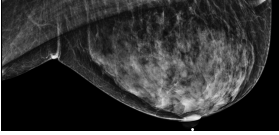
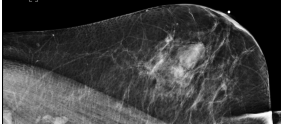
Correct cases (MM-11387, MM-9467): In BI-RADS density classification and benign/malignant discrimination, GPT-5 successfully identified classic radiological features. In the correct density example, the model accurately recognized an almost entirely dense breast. In the correct malignancy example, the model correctly identified a lesion with an irregular spiculated mass, architectural distortion, and suspicious calcifications (hallmark features of malignancy).

Error cases (MM-4307, MM-20757): For density assessment errors, the model consistently misclassified BI-RADS category D (extremely dense) as category C (heterogeneously dense), indicating a tendency to underestimate overall density in highly dense breasts. For malignancy assessment errors, benign structural changes were misclassified as malignant (overdiagnosis), suggesting the model is prone to false-positive interpretations when confronted with architectural distortion or irregular mass margins lacking other corroborating malignant features.

4 Discussion

Our evaluation indicates that ChatGPT-5, as a general-purpose multimodal LLM, struggles with complex, domain-specific queries in mammography. This limitation is likely due to the model’s lack of fine-tuning on high-resolution

Table 6: Representative GPT-5 predictions: two correct and two incorrect cases in BI-RADS density and malignancy classification.

| Correct: (Yes) | Malignancy | Correct: (Yes) | Malignancy | Error: Density D→C | Error: B→A | Malignancy |
|---|------------|--|------------|--|---------------|---|
|  | |  | |  | |  |
| GT: Yes Pred: Yes Rationale: The mammogram does not show a discrete spiculated mass, suspicious pleomorphic/linear branching calcifications, or convincing architectural distortion. Findings appear consistent with normal fibroglandular tissue without specific features of malignancy. | | GT: Yes Pred: Yes Rationale: (A) Yes. The image shows an irregular spiculated mass with architectural distortion and suspicious calcifications—features suggestive of malignancy. | | GT: D Pred: C Rationale: The breast shows large regions of fibroglandular tissue with some interspersed fat—heterogeneously dense, not almost entirely fatty (A/B) nor uniformly/extremely dense (D). | | GT: No Pred: Yes Rationale: (A) Yes. The mass is irregular and spiculated with associated architectural distortion—features that are highly suspicious for malignancy. |

grayscale breast imaging and the absence of explicit adaptation to structured clinical tasks such as BI-RADS classification or lesion characterization. Mammogram interpretation itself is inherently challenging, even for experienced radiologists. Diagnostic decisions often hinge on subtle, low-contrast features, such as microcalcification clusters, spiculated mass margins, or architectural distortions that can be obscured by overlapping fibroglandular tissue, particularly in dense breasts, and further complicated by variations in acquisition parameters and image quality. Accurate assessment requires not only fine-grained visual discrimination but also the integration of patient history, comparison with prior examinations, and adherence to standardized reporting frameworks like BI-RADS. These factors underscore why mammographic VQA is especially difficult for general-purpose AI systems without targeted domain adaptation and large-scale, modality-specific training.

4.1 Limitations

This study has several limitations. First, both the evaluation data and the comparison model outputs were drawn from the same datasets, but a fully unified test set across all models was not feasible. For datasets with extremely large sample sizes, it was impractical to perform exhaustive evaluation on the entire test split. Instead, we adopted a balanced sampling strategy, evenly distributing samples across all question types and answer categories to ensure representative coverage. As such, the reported results should be regarded as indicative rather than exhaustive. Second, all experiments were conducted in a zero-shot setting using ChatGPT-5 without domain-specific fine-tuning or prompt optimization for mammography. This design allows us to assess the model’s intrinsic, out-of-the-box capabilities, but may underestimate the achievable performance with targeted adaptation. Third, the evaluation was limited to a fixed set of question templates derived from structured labels, which may not capture the full variability and nuance of real-world clinical questioning. Fourth, while several SOTA methods have reported strong results on the same datasets, direct comparison remains challenging due to differences in experimental setups, data preprocessing pipelines, and evaluation protocols. Finally, our analysis focused primarily on accuracy-based metrics; other important aspects such as reasoning transparency, uncertainty calibration, and systematic error characterization were not explored.

5 Conclusion

In this study, we preliminarily evaluated the zero-shot performance of the ChatGPT-5 family and GPT-4o model on clinically relevant VQA tasks in digital and digitized mammography, using four diverse public datasets (EMBED, InBreast, CMMD, and CBIS-DDSM). ChatGPT-5 lags behind physicians and specialized, fine-tuned SOTA models on most tasks and is prone to false-positive interpretations when confronted with architectural distortion or irregular mass margins. While not ready for clinical use, GPT-5 achieved considerable improvement over the GPT-4o model,

so we are optimistic about the potential for generalist multimodal LLMs to improve patient care in mammography VQA tasks.

Code Availability

The code used in this study is publicly available at [GPT-5-Evaluation](#).

References

- [1] Joanne Kim, Andrew Harper, Valerie McCormack, Hyuna Sung, Nehmat Houssami, Eileen Morgan, Miriam Mutebi, Gail Garvey, Isabelle Soerjomataram, and Miranda M Fidler-Benaoudia. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine*, pages 1–9, 2025.
- [2] Constance D Lehman, Robert F Arao, Brian L Sprague, Janie M Lee, Diana SM Buist, Karla Kerlikowske, Louise M Henderson, Tracy Onega, Anna NA Tosteson, Garth H Rauscher, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49–58, 2017.
- [3] Fiona J. Gilbert, Susan M. Astley, Michael A. McGee, Mary G. Gillan, Christopher R. Boggis, Paul M. Griffiths, and Stephen W. Duffy. Single reading with computer-aided detection and double reading of screening mammograms in the united kingdom national breast screening program. *Radiology*, 241(1):47–53, October 2006.
- [4] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.
- [5] Yan Chen, Adnan G Taib, Iain T Darker, and Jonathan J James. Performance of a breast cancer detection ai algorithm using the personal performance in mammographic screening scheme. *Radiology*, 308(3):e223299, 2023.
- [6] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists’ performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019.
- [7] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer methods and programs in biomedicine*, 157:19–30, 2018.
- [8] Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.
- [9] Mojtaba Safari, Shansong Wang, Mingzhe Hu, Zach Eidex, Qiang Li, and Xiaofeng Yang. Performance of gpt-5 in brain tumor mri reasoning, 2025.
- [10] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
- [11] Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*, 2025.
- [12] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [13] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- [14] Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilie, Geoffrey Smith, et al. The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1):e220047, 2023.
- [15] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.
- [16] Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, and Jinghua Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast. *The Cancer Imaging Archive*, 1, 2021.

- [17] Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). (No Title), 2016.
- [18] OpenAI. Gpt-5. <https://openai.com/research>, 2025. Accessed: 2025-08-14.
- [19] Xuxin Chen, Yuheng Li, Mingzhe Hu, Ella Salari, Xiaoqian Chen, Richard LJ Qiu, Bin Zheng, and Xiaofeng Yang. Mammo-clip: Leveraging contrastive language-image pre-training (clip) for enhanced breast cancer diagnosis with multi-view mammography. arXiv preprint arXiv:2404.15946, 2024.
- [20] Luhao Sun, Bowen Han, Wenzong Jiang, Weifeng Liu, Baodi Liu, Dapeng Tao, Zhiyong Yu, and Chao Li. Multi-scale region selection network in deep features for full-field mammogram classification. Medical Image Analysis, 100:103399, 2025.
- [21] Han Chen and Anne L Martel. Enhancing breast cancer detection on screening mammogram using self-supervised learning and a hybrid deep model of swin transformer and convolutional neural networks. Journal of Medical Imaging, 12(S2):S22007–S22007, 2025.
- [22] Naga Raju Gudhe, Sudah Mazen, Reijo Sund, Veli-Matti Kosma, Hamid Behravan, and Arto Mannerman. A multi-view deep evidential learning approach for mammogram density classification. IEEE Access, 12:67889–67909, 2024.
- [23] Eleonora Lopez, Eleonora Grassucci, Martina Valleriani, and Danilo Comminiello. Multi-view hypercomplex learning for breast cancer screening. arXiv preprint arXiv:2204.05798, 2022.
- [24] Tao Wei, Angelica I Aviles-Rivero, Shuo Wang, Yuan Huang, Fiona J Gilbert, Carola-Bibiane Schönlieb, and Chang Wen Chen. Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations. Medical image analysis, 82:102618, 2022.